

Categorization and Reference-Dependent Decision-Making under Statutory Thresholds: Evidence from a Criminal Law Reform

March 9, 2026

Tereza Burýšková¹²

Abstract

Legal and regulatory regimes frequently organize decisions through statutory categories defined by quantitative thresholds. I exploit a criminal law reform that shifted statutory damage thresholds, creating a natural experiment that allows me to study how changes in categorical boundaries affect discretionary decision-making. The empirical strategy exploits differential exposure to the reform across offense categories within a difference-in-differences framework. I find that judges impose significantly shorter prison sentences both when statutory ranges are explicitly reduced and when the formal statutory range remains unchanged. The effects arise primarily on the intensive margin of sentencing, with little change in the probability of imprisonment. This evidence is consistent with two mechanisms: a severity channel, whereby statutory categories signal the expected level of punishment, and a reference channel, whereby decision-makers evaluate cases relative to others within the same category. More broadly, the results suggest that reforms altering categorical boundaries can generate spillover effects through changes in reference groups, even in the absence of formal changes to applicable rules. The findings have implications both for the design of threshold-based policies and for empirical strategies that treat formally unaffected categories as valid control groups.

JEL classification: K14, K42, D91

Keywords: categorization; reference dependence; statutory thresholds; judicial decision-making; institutional design

¹Faculty of Law, Charles University, náměstí Curieových 901/7 116 40 Prague 1, Czech Republic, tereza.buryskova164@student.cuni.cz

²I thank Keren Weinshall, Michal Šoltés, and Jakub Drápal for their helpful comments. The project has received funding from the European Regional Development Fund (Center for Inequality and Open Society, Project No. CZ.02.01.01/00/23_025/0008690)

1 Introduction

Many legal and regulatory regimes allocate decisions across statutory categories defined by quantitative thresholds. In criminal law, taxation, and regulatory policy, such thresholds determine which rules apply and often separate qualitatively different legal regimes. Typically, the distinction relies on observable characteristics (e.g., monetary damage, offense type), creating discrete decision environments. While designed as formal constraints, these categories may also influence decision-making through informational and reference-based channels, generating behavioral spillovers beyond their formal scope.

This paper provides evidence that statutory thresholds affect judicial decisions not only through formal changes in sentencing ranges, but also through changes in the reference group within which cases are evaluated. I exploit a 2020 criminal law reform that adjusted monetary thresholds for theft offenses, creating exogenous variation in category membership. The reform reclassified certain theft offenses by adjusting monetary thresholds, thereby altering the composition of cases within statutory categories without uniformly changing the applicable ranges.

The evidence is consistent with two distinct and empirically separable mechanisms. Judicial responses arise from both direct changes in expected punishment induced by formal range adjustments (a severity channel) and shifts in evaluation driven by changes in the reference group within a statutory category (a reference channel). Importantly, this implies that threshold reforms can influence decisions even in categories whose formal legal rules remain unchanged, because the reference group within which cases are evaluated shifts.

The severity channel reflects the primary rationale for statutory ranges: they provide a discrete classification of offenses that signals legislatively intended punishment levels. Accordingly, cases assigned to a lower statutory range should, all else equal, receive lower sentences, even if the originally imposed sentence remains legally admissible.

The reference channel arises because decision-makers evaluate cases relative to others within the same statutory category, forming endogenous reference groups within fixed formal constraints. When more severe cases are added to a category without changing the formal statutory range, the distribution of reference cases shifts within the category, potentially lowering sanctions for previously typical cases. Reference-dependent evaluation has been extensively studied in behavioral economics and psychology (Kőszegi & Rabin, 2006; Bordalo et al., 2012) and has been applied to legal decision-making contexts (Englich et al., 2006; Leibovitch, 2017), providing a microfoundation of the mechanisms examined here. More broadly, the results highlight how threshold-based institutional design can generate behavioral spillovers that are not captured by formal rule changes alone.

This paper relates most closely to Drápal and Šoltés (2024), who develop a theoretical and

experimental framework distinguishing between a severity channel—where statutory ranges signal legislatively intended punishment levels—and a reference channel—where cases are evaluated relative to others within the same statutory category. While their analysis relies on controlled experimental variation, the present paper provides complementary evidence from a real-world institutional setting. By exploiting a nationwide reform of statutory thresholds in the Czech Criminal Code, I study how these mechanisms operate in actual judicial decision-making using observational case-level data. Moreover, the institutional reform creates variation not only in statutory ranges but also in the composition of cases within categories, allowing me to document spillover effects across formally unchanged categories. This evidence highlights how threshold reforms can reshape decision environments beyond their direct legal scope.

The empirical design exploits differential exposure to the reform across offense categories. Using a difference-in-differences framework, I find that judicial responses arise primarily on the intensive margin, through reductions in sentence length. These reductions occur in both cases with lowered sentencing ranges and cases whose ranges remained unchanged but were expanded to include more severe offenses. The pattern of results is consistent with a severity channel for directly affected categories and a reference channel for categories whose formal ranges remained unchanged but whose composition shifted.

Conversely, the reform had little effect on the extensive margin of sentencing. The probability of custodial imprisonment remains largely unchanged. This stability sharpens the interpretation of the intensive-margin effects: judges did not switch between punishment types but instead adjusted sentence lengths within the existing framework of custodial punishment.

By disentangling severity and reference effects within a threshold-based regime, this paper shows how formally neutral reforms can systematically reshape decision-making environments in empirically identifiable ways. More broadly, the findings highlight a challenge for empirical designs evaluating threshold policies: categories that appear formally untreated may still respond through reference-group spillovers.

The remainder of the paper proceeds as follows. Section 2 reviews the related literature. Section 3 describes the institutional setting and the 2020 reform. Section 4 presents the data and empirical strategy. Section 5 reports the main results and robustness checks. The final section concludes.

2 Literature Review

This paper relates to three strands of literature. First, it contributes to the economics of threshold-based institutional design, which studies how quantitative cutoffs structure incentives and behavior. Second, it connects to research on reference-dependent evaluation and catego-

rization, which emphasizes that decisions depend on relative position within salient comparison groups. Third, it adds to the empirical literature on judicial decision-making and sentencing discretion. By combining these strands, the paper highlights how shifts in statutory categories can generate both direct and indirect behavioral responses, even when formal legal constraints remain unchanged.

Threshold-Based Institutional Design

A large body of economic literature examines how quantitative thresholds structure incentives and behavior in regulatory and institutional environments. Thresholds generate discrete regimes in otherwise continuous settings and often induce discontinuous behavioral responses around policy cutoffs (Lee, 2008; Saez, 2010). Much of this literature focuses on behavioral responses at the margin of the cutoff itself—for example, bunching around tax thresholds or regulatory limits.

Less attention has been paid to how reforms that shift threshold levels reshape the composition of observations within categories defined by those thresholds. When category boundaries change, units that were previously classified in one regime may be reassigned to another, potentially altering the distribution of cases within each category. These compositional changes may affect behavior even for actors whose formal constraints remain unchanged. Understanding such effects is therefore important for evaluating the broader behavioral consequences of threshold-based institutional design.

Reference Dependence and Categorization

A complementary literature in behavioral economics emphasizes that evaluations are reference-dependent and shaped by salient comparisons (Kőszegi & Rabin, 2006; Bordalo et al., 2012). Under reference-dependent preferences, outcomes are assessed relative to a reference point that is often determined by contextual information or salient comparison groups. When decisions occur within institutional categories, the relevant reference point may depend on the distribution of cases within that category.

Related work in behavioral law and economics highlights how legal judgments may exhibit context dependence. For example, experimental evidence shows that legal evaluations such as damage awards can vary depending on the set of comparison cases presented to decision-makers (Kelman, Rottenstreich, & Tversky, 1996). These findings suggest that decision-makers may evaluate cases relative to other available cases rather than solely according to their absolute characteristics. However, empirical evidence on reference-dependent behavior among professional decision-makers operating within real-world institutional constraints remains relatively limited.

Sentencing as a Discretionary Institutional Setting

Criminal sentencing provides a natural environment in which formal legal rules coexist with substantial discretion. Empirical research documents that sentencing outcomes vary across judges and institutional settings even when statutory guidelines are present (Anderson, Kling, & Stith,

1999; Ulmer & Johnson, 2004). Reforms to sentencing guidelines and mandatory minimum statutes have also been shown to influence punishment outcomes, particularly when statutory constraints directly restrict judicial discretion (Bjerk, 2017). Additional evidence suggests that sentencing decisions may respond to institutional factors such as enforcement practices and legal thresholds (Skugarevskiy, 2017).

A related literature documents substantial heterogeneity in decision-making across judges even within the same institutional framework. Using administrative data from U.S. courts Abrams, Bertrand, and Mullainathan (2012) document substantial heterogeneity in decision-making across judges even within the same institutional framework. Using administrative data from U.S. courts, they show that judges differ systematically in their sentencing behavior, highlighting the importance of accounting for judge-level variation when studying judicial outcomes. Similarly, (Anwar & Fang, 2015) study systematic variation in judicial decisions across judges in criminal bail determinations, further illustrating how individual decision-makers can influence legal outcomes even under shared institutional rules.

Recent work further highlights that sentencing outcomes may depend on comparative and contextual factors. In particular, (Leibovitch, 2016, 2017) shows that punishments may depend on the relative severity of cases within a judge's docket rather than solely on the absolute characteristics of an offense. From this perspective, judges implicitly evaluate cases within a distribution of comparable cases, creating systematic relative-judgment effects in sentencing outcomes. (Danziger, Levav, & Avnaim-Pesso, 2011) shows that parole decisions fluctuate systematically over the course of a judge's decision session, suggesting that even professional legal decision-makers may rely on heuristics when evaluating cases under time pressure. These findings highlight how features of the decision environment can shape legal outcomes alongside formal legal rules.

Taken together, this literature suggests that judicial decisions are shaped not only by formal legal rules but also by the institutional and informational environments in which judges operate.

Closest to this paper, Drápal and Šoltés (2024) analyze sentencing decisions around statutory quantity thresholds and identify two mechanisms shaping punishment decisions: a severity channel, in which statutory sentencing ranges signal legislatively intended punishment levels, and a reference channel, in which cases are evaluated relative to others within the same statutory category. Their analysis relies on experimental variation designed to isolate these mechanisms in a controlled environment.

This paper complements that work by providing field evidence from an institutional reform that reclassified offenses through changes in statutory thresholds. Using observational sentencing data, I examine how these mechanisms operate in a real-world judicial environment. In addition to identifying direct responses to changes in statutory sentencing ranges, the analysis shows that reforms can generate spillover effects across categories whose formal sentencing ranges remain unchanged.

These findings extend the experimental evidence by demonstrating how threshold reforms reshape decision environments through changes in both formal legal constraints and reference groups.

Institutional Design of Courts

More broadly, the paper contributes to the literature studying how institutional design shapes judicial decision-making. A growing body of empirical work emphasizes that legal outcomes depend not only on statutory rules but also on institutional features such as case assignment mechanisms, docket composition, and procedural structures. For example, (Chen, Moskowitz, & Shue, 2016) show that decision-makers in legal and quasi-legal settings may exhibit systematic behavioral patterns driven by the sequence of cases they encounter.

By focusing on statutory thresholds as a design feature of legal institutions, this paper highlights how categorical legal rules structure the informational environment in which discretionary decisions are made. Changes in statutory categories can therefore influence judicial outcomes not only by altering formal sentencing ranges but also by redefining the set of cases that serve as implicit reference points for judicial evaluation.

3 Institutional Context

3.1 Sentencing Structure and Judicial Discretion

In the Czech Republic, criminal sentencing combines rigid statutory boundaries with substantial intra-range judicial discretion. Every offense defined in the special part of the Criminal Code is assigned a statutory sentencing range consisting of a lower and an upper bound of imprisonment.

Criminal cases are assigned to judges according to a predetermined allocation schedule. The assignment mechanism operates independently of the statutory damage thresholds studied here. As a result, conditional on filing, theft cases are effectively randomly distributed across judges within a court. This institutional feature limits strategic case sorting and ensures that changes in sentencing patterns following the reform cannot be attributed to systematic reassignment of case types to particular judges.

Judges are required to impose a sentence within this range in nearly all cases, which makes statutory category assignment legally consequential. Within the applicable range, however, judges retain broad discretion and are not constrained by formal sentencing guidelines or centralized sentencing recommendations. Within these statutory bounds, judges retain broad discretion. The Czech system does not operate under binding sentencing guidelines (apart from the ranges themselves), point systems, or sentencing commissions, and no centralized body issues quantitative within-range recommendations. Sentencing practice is therefore shaped primarily by individual judicial judgment rather than coordinated institutional benchmarks.

In standard theft cases without aggravating circumstances, statutory classification is mechanically determined by the monetary value of damage. Damage is defined as the objective market value of the stolen property at the time of the offense. For these cases, no additional qualitative assessment affects category assignment. I therefore restrict the analysis to standard theft cases in which statutory categorization is fully determined by the amount of monetary damage.

Apart from the custodial sentence, judges can also use other types of punishment (suspended sentence, monetary sanctions, etc.). While the type of punishment is discretionary, the length of custodial sentences must fall within the applicable statutory range. The reform studied here affects statutory classification but does not alter the menu of available sanctions.

3.2 The 2020 Threshold Reform

In October 2020, a reform of the Czech Criminal Code increased nominal monetary thresholds that define damage categories. The reform adjusted the quantitative definitions of damage without altering sentencing principles, judicial discretion within ranges, or procedural rules. The structure of statutory sentencing ranges remained unchanged; only the monetary cutoffs determining category membership shifted upward. For example, a case involving CZK 60k damage was subject to a statutory range of 1–5 years' imprisonment. After the reform, the same nominal damage was reclassified into a lower category, with a range of 0–2 years.

Importantly, for categories whose formal statutory range remained unchanged, the reform altered only the composition of cases within the category. Judges retained identical formal discretion before and after the reform. Importantly, for categories whose formal statutory range remained unchanged, the reform altered only the composition of cases within the category. Judges retained identical formal discretion before and after the reform.

Table 1 summarises the legal design of the reform. This design implies two distinct types of variation. For some damage ranges, the sentencing range shifted, and at the same time, they were grouped with a less severe group of cases. For instance, cases with damage of CZK 50k-100k shifted from 1-5 years to 0-2 years and got grouped with 10k-50k. I denote this type of variation as Treatment A. Other damage ranges kept their original statutory range, but that range expanded to include cases with more serious monetary damage. For instance, cases with damage of CZK 100k-500k face the same sentencing range of 1-5 years; however, the same sentencing range now captures a damage range of CZK 100k-1m, which is relatively more severe than the original damage range of CZK 50k-500k.

Table 2 highlights the distinction between Treatment A and Treatment B and relates it to my research design. The black box shows the samples used in the main analysis

Table 1: Sentencing Ranges for Theft Cases in the Criminal Code

Damage (CZK)	Sentencing Range	
	Before October 2020	Starting October 2020
less than 5k	not a criminal offense	not a criminal offense
5k-10k	0-2 years	
10k-50k		0-2 years
50k-100k	1-5 years	
100k-500k		1-5 years
500k-1m	2-8 years	
1m-5m		2-8 years
5m-10m	5-10 years	
more than 10m		5-10 years

Note: The sentencing ranges for different types of theft as prescribed by the Criminal Code before and after the 2020 reform. The reform came into power on October 1, 2020.

4 Empirical Strategy

4.1 Data

In the Czech Republic, criminal cases are comprehensively documented, and detailed case-level data are available for research. In particular, there is data about the criminal procedure, including the court and judge identifier; second, the data about the offense, mainly its legal classification and corresponding section and paragraph in the Criminal Code, and the damage caused, where relevant, and data about the defendant (ethnicity, gender, etc.). Since this data is directly reported by the court officers and captures the evaluation of all evidence presented, it should be of sufficient quality.

The data span the period from 2006 to 2023. However, the damage caused, which is central to my analysis, has been reported only for cases decided after October 2019, 12 months before the reform. Appendix Figure A.1 shows that a stable report rate of around 50 % emerged by the beginning of 2020, as the court staff started to adjust to the newly introduced variables. As a result, the pre-reform period used in the empirical analysis is relatively short. This limitation primarily affects the ability to test long-run pre-trends. However, the reform was implemented abruptly and was not anticipated when the dataset was constructed. Moreover, the event-study estimates reported below show that the treated and control groups exhibit parallel trends in sentencing out-

Table 2: Two Different Types of Theft Cases in Terms of Reform Effects

Damage (CZK)	Sentencing Range		Sentencing Range Thresholds	Damage Range Thresholds
	Before Reform	After Reform		
Treatment A				
5k–10k	0–2 years	not a criminal offense	↓	↑
50k–100k	1–5 years	0–2 years	↓	↑
500k–1m	2–8 years	1–5 years	↓	↑
5m–10m	5–10 years	2–8 years	↓	↑
Treatment B				
less than 5k	not a criminal offense	not a criminal offense	–	↑
10k–50k	0–2 years	0–2 years	–	↑
100k–500k	1–5 years	1–5 years	–	↑
1m–5m	2–8 years	2–8 years	–	↑
more than 10 m	5-10 years	5-10 years	–	↑

Note: A distinction between Treatment A and Treatment B type cases based on the impact of the 2020 reform. Own summary based on the Criminal Code.

comes during the available pre-reform period, providing support for the identifying assumptions of the difference-in-differences design.

Appendix Table A.1 presents the descriptive statistics of the dataset.

4.2 Mechanism Identification

Statutory sentencing ranges may affect judicial decisions through two distinct mechanisms. First, ranges function as categorical signals of legislatively intended punishment levels. Assignment to a lower statutory range reduces the formal benchmark against which punishment is determined (severity channel). Second, ranges define comparison sets within which cases are evaluated relative to one another. Changes in the composition of cases within a range may therefore shift sentences even when statutory bounds remain unchanged (reference channel).

The 2020 reform generates institutional variation, allowing these mechanisms to be separated.

Treatment B: Isolating the Reference Channel

In Treatment B categories, statutory sentencing ranges remain unchanged, but the distribution of damage expands upward. Because the formal punishment benchmark is constant, the severity

channel predicts no change in sentencing. However, including more severe cases shifts the internal distribution of damage within the category. Previously, typical cases occupied a relatively lower position within their reference group. Under reference-dependent evaluation, this compositional shift results in shorter sentences.

Prediction 1 (Reference Effect). Under reference effect, in categories whose statutory sentencing range remains unchanged, the reform decreases mean sentence length.

Treatment A: Competing Channels

Treatment A induces a twofold variation: the cases become subject to a lower statutory range and, at the same time, relatively more severe within the new category. The severity channel predicts lower sentences due to the reduced formal punishment benchmark. At the same time, cases with less severe offenses are pooled into the new range, potentially moving them upward in the within-category distribution. Under reference-dependent evaluation, this relative repositioning increases the number of sentences.

Prediction 2 (Competing Effects). In categories experiencing a downward shift in statutory sentencing ranges, the observed change in mean sentence length reflects the relative magnitude of severity and reference effects. A decrease implies dominance of the severity channel; an increase implies dominance of the reference channel.

Table 3: A Summary of the Impact of Severity and Reference channel on the Sentences

	Severity	Reference
	Impact on Mean Sentence	
Treatment A (<i>Sentencing Range Downward Shift</i>)	↓	↑
Treatment B (<i>Damage Range Upward Shift</i>)	0	↓

Note: A summary of the impact of severity and reference channel (as introduced by Drápal and Šoltés (2024)) on cases analyzed in this paper. The severity channel decreases the sentence for Treatment A cases and does not influence Treatment B cases. The reference channel increases sentences for Treatment A and decreases them for Treatment B. This intuition provides an interpretation of the empirically estimated changes in average sentence length.

4.3 Estimation

To ensure comparability of treated and control units, I restrict the treated theft sample to two damage intervals directly affected by the 2020 threshold reform: (i) CZK 50k–100k (Treatment A) and (ii) CZK 10k–50k (Treatment B).

I further limit the estimation window to event times $\tau \in [-6, 8]$ quarters around the reform, where $\tau = 0$ denotes the reform quarter (Q3 2020).

I study two sentencing outcomes for non-suspended custodial sentences. First, the extensive margin:

$$Y_{ijt} = \mathbf{1}\{\text{custodial imprisonment}\}.$$

Second, the intensive margin:

$$Y_{ijt} = \text{length of custodial imprisonment in months}.$$

4.3.1 Difference-in-Differences

I estimate a pooled difference-in-differences specification including Treatment A, Treatment B, and control cases. Let $After_t$ be an indicator for decisions issued after September 30, 2020. Let $\mathbf{1}\{G_i = A\}$ and $\mathbf{1}\{G_i = B\}$ denote membership in the two treated groups, with the control group serving as the omitted category.

The baseline specification is:

$$Y_{ijt} = \beta_A (After_t \times \mathbf{1}\{G_i = A\}) + \beta_B (After_t \times \mathbf{1}\{G_i = B\}) + \gamma_j + \lambda_t + X'_{ijt} \delta + \varepsilon_{ijt}, \quad (1)$$

where γ_j denotes judge fixed effects and λ_t denotes calendar-quarter fixed effects, X_{ijt} includes indicators for recidivism, age, gender, and concurrence. Identification relies on within-judge variation over time, netting out common time shocks using the control offenses.

Including judge-fixed effects is important in this setting because sentencing practices vary substantially across judges. These fixed effects ensure that identification relies on within-judge variation over time, comparing how the same judge sentences comparable cases before and after the reform. This absorbs time-invariant differences in sentencing severity across judges and isolates changes in sentencing behavior induced by the reform rather than differences in the composition of judges deciding particular cases

Under Prediction 1, the coefficient β_B identifies the reference effect for Treatment B. Under Prediction 2, the coefficient β_A reflects the net effect of competing severity and reference channels for Treatment A.

4.3.2 Event-study

To assess pre-trends and dynamic effects, I estimate an event-study version of the model. Let $D_{\tau(i)}$ denote an indicator for event time τ , with $\tau = -1$ as the omitted reference period. All other notation remains the same. The specification is:

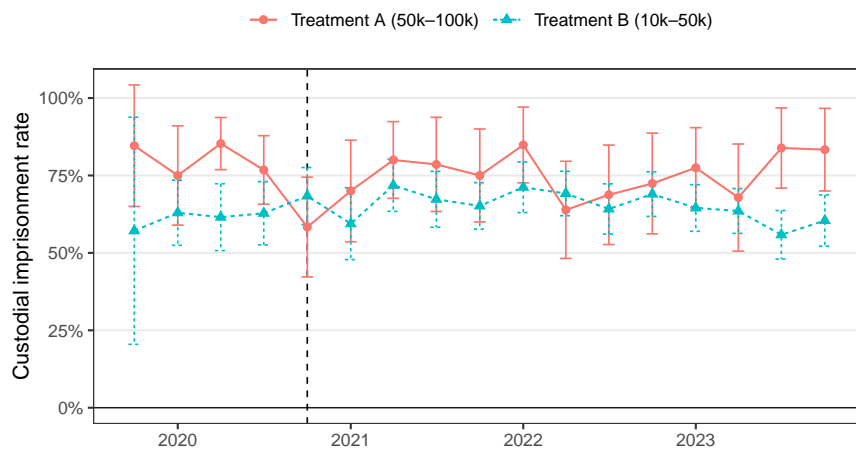
$$Y_{ijt} = \sum_{\tau \neq -1} [\beta_{\tau,A} (D_{\tau(i)} \times \mathbf{1}\{G_i = A\}) + \beta_{\tau,B} (D_{\tau(i)} \times \mathbf{1}\{G_i = B\})] + \gamma_j + \lambda_t + X'_{ijt} \delta + \varepsilon_{ijt}. \quad (2)$$

The coefficients $\beta_{\tau,A}$ and $\beta_{\tau,B}$ measure treatment effects relative to the last pre-reform quarter. Because the applicable legal framework is determined by the date on which the sentence takes legal effect, the treatment indicator is defined relative to October 1, 2020, when the reform entered into force.

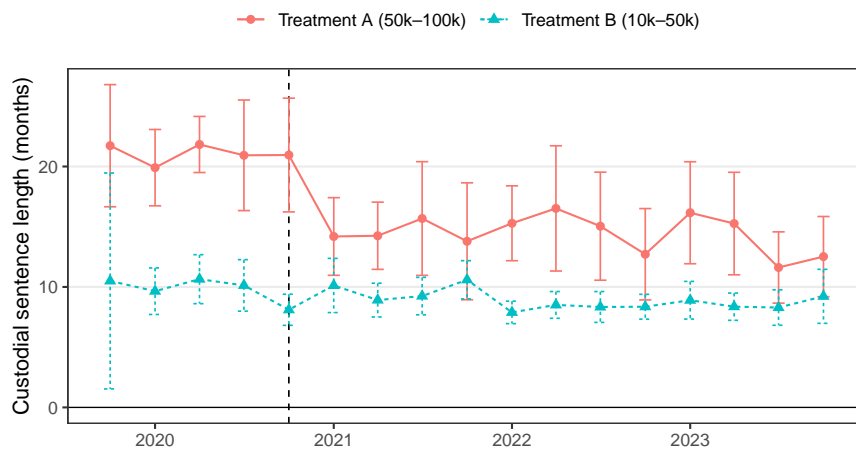
5 Results

5.1 Descriptive Evidence

Figure 1 plots the quarterly evolution of custodial sentence rates and sentence lengths for the two treated groups. While the probability of custodial imprisonment remains broadly stable, the average custodial sentence length declines sharply for Treatment A immediately after the reform. For Treatment B, the decline appears less clear.



(a) Extensive margin



(b) Intensive margin

Figure 1: The time evolution of custodial sentence probability and length

Note: Panel (a) shows the custodial imprisonment rate, panel (b) the custodial imprisonment length. 95 percent confidence intervals are shown. The black dashed line denotes the adoption of the reform.

5.2 Baseline DD estimates

Given the scope of the 2020 reform, it is challenging to identify a set of control cases that were completely unaffected. For instance, all crimes against property were at least partially affected, which unfortunately prevents them from serving as a control group in my analysis.

For the main analysis, I use negligence of mandatory support (§196) as the control offense. This crime is regularly adjudicated by the same criminal courts and judges that decide theft cases, and sentencing is determined under the same institutional framework of statutory ranges and judicial discretion. At the same time, the legal classification of this offense does not depend on the monetary damage thresholds modified by the 2020 reform. Consequently, the reform did not alter either the statutory sentencing ranges or the criteria determining category membership for this offense. This makes it a suitable benchmark for capturing common time shocks in sentencing behavior unrelated to the reform.

Table 4 reports pooled DD estimates for both the extensive and intensive margins of custodial sentencing. The coefficients of interest are the interaction terms between the post-reform indicator and the treatment group dummies.

Panel A shows that the reform does not generate a statistically robust change in the probability of imposing a custodial sentence. Only with Treatment A are the coefficients positive at the 10 % level; however, their magnitude and significance decrease after accounting for the fixed effects. In Appendix Figure A.2, I show that this is mostly driven by a narrow range of thefts between 70k-80k. Therefore, this tiny glimpse in the extensive margin should not impair the results for the intensive margin.

Panel B reveals a clear, economically meaningful decline in the length of custodial sentences. For Treatment A, the estimated reduction is around 5 months (depending on the specification) and remains highly statistically significant after including judge and quarter fixed effects and defendant controls. Given the pre-reform average sentence length of 21.1, this represents a substantial decrease in punishment severity within this group.

For Treatment B, the estimated decline is smaller, around one month. Three out of four specifications are significant at least at the 10 % level (p-values 0.11, 0.07, 0.08, 0.04 respectively). The richest specification delivers the strongest significance. Moreover, the point estimates remain similar, further supporting the interpretation of the reform's effect.

Table 4: Baseline Difference-in-Differences Estimates

	(1)	(2)	(3)	(4)
Panel A. Extensive margin: Custodial imprisonment (LPM)				
Treatment A × After	-0.078+ (0.045)	-0.038 (0.044)	-0.078+ (0.045)	-0.043 (0.044)
Treatment B × After	0.049 (0.037)	0.046 (0.040)	0.052 (0.038)	0.047 (0.040)
Observations	20,947	20,947	20,947	20,947
R^2	0.006	0.056	0.010	0.060
Controls			✓	✓
Judge FE		✓		✓
Quarter FE		✓		✓
Panel B. Intensive margin: Custodial sentence length (months)				
Treatment A × After	-5.764*** (1.485)	-5.832*** (1.465)	-4.891*** (1.348)	-5.142*** (1.373)
Treatment B × After	-1.251 (0.775)	-1.365+ (0.744)	-1.188+ (0.674)	-1.297* (0.639)
Observations	12,046	12,046	12,046	12,046
R^2	0.039	0.103	0.055	0.115
Controls			✓	✓
Judge FE		✓		✓
Quarter FE		✓		✓

Notes: The table reports pooled difference-in-differences estimates comparing Treatment A and Treatment B cases to a control group. Standard errors (in parentheses) are clustered at the judge level. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Controls include recidivism, offender age and gender, and a concurrence indicator.

5.3 Dynamic effects and pre-trends

Figure 2 complements the pooled DD estimates by reporting dynamic difference-in-differences coefficients relative to the control group and normalized to the last pre-reform quarter ($t = -1$) tracing the quarter-specific evolution of treatment effects before and after the reform.

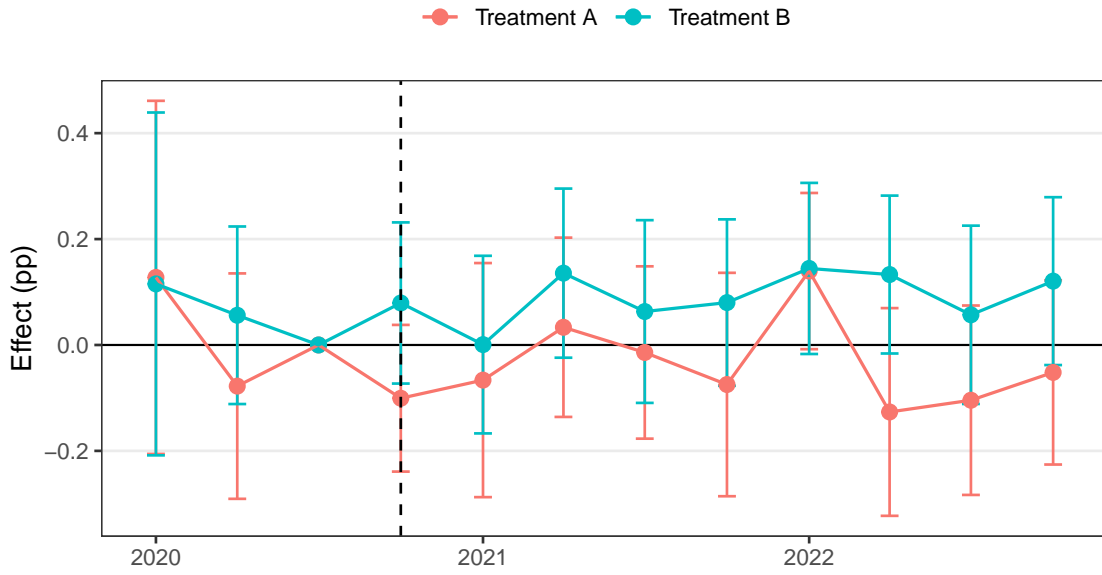
Importantly, the event-study estimates show no evidence of differential pre-reform trends between treated theft cases and the control offense, supporting the validity of the parallel trends assumption.

Turning to the extensive margin (Panel (a)), the dynamic estimates confirm the absence of a robust response in the probability of custodial sentencing. For Treatment A, the post-reform coefficients are slightly negative in several quarters, while for Treatment B, they are close to zero and occasionally mildly positive. However, the estimates are small in magnitude, indicating that the reform did not generate a systematic shift in the extensive margin of incarceration.

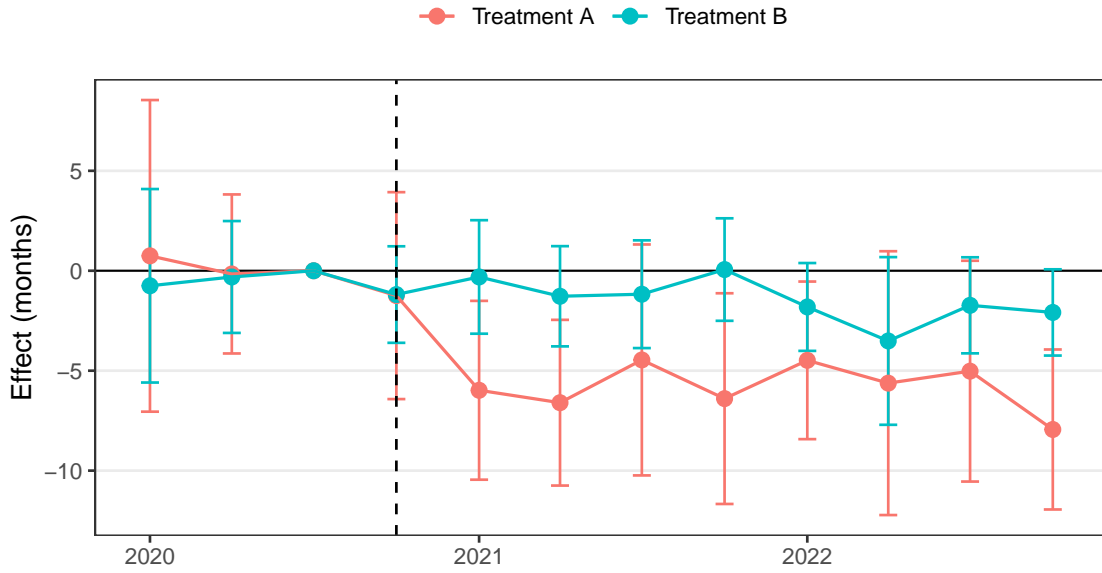
In contrast, Panel (b) shows a clear and persistent adjustment along the intensive margin. For Treatment A, sentence lengths decline sharply immediately after the reform and remain substantially below the pre-reform benchmark relative to the control group. The post-reform effects are economically large and stable over time, consistent with the pooled DD estimates indicating a reduction of approximately 4.5 to 5.9 months. Given the pre-reform average of 21.1 months, this represents a sizeable decrease in punishment severity.

For Treatment B, the post-reform decline in sentence length is considerably smaller in magnitude. The dynamic coefficients are generally negative but remain closer to zero and lack statistical power. This pattern aligns with the baseline estimates, which indicate a modest reduction of roughly one month relative to a pre-reform mean of 10.2 months.

Overall, the dynamic evidence reinforces the conclusion that the reform operated primarily along the intensive margin. The immediate and sustained decline for Treatment A is consistent with a mechanical severity effect induced by the downward shift in the statutory sentencing range. The smaller and more gradual adjustment for Treatment B is consistent with a weaker response within an unchanged formal range, potentially reflecting gradual adaptation to the expanded case mix.



(a) quarterly effects on custodial sentence rate (extensive margin)



(b) quarterly effects on custodial sentence length (intensive margin)

Figure 2: Dynamic DD for custodial sentence

5.4 Robustness checks

This section summarizes a series of robustness checks designed to assess the stability of the main difference-in-differences estimates. Detailed results are reported in the Appendix.

- Appendix A3 examines the distribution of punishment types and shows that it remains largely unchanged, supporting the dominance of the intensive margin.
- Appendix A4 investigates the distributional effects of the reform, plotting the sentence as a function of damage. It demonstrates that the reform affected the entire distribution of cases.
- Appendix A5 introduces an alternative control group and shows that the estimates are similar to those in the main analysis, indicating they are not control-group dependent.
- Appendix A6 constructs broader treatment groups by pooling damage intervals affected by the reform in a similar way. The resulting estimates remain qualitatively consistent with the baseline results.

5.5 Interpretation and Mechanisms

The empirical results are summarized in Table 5. Across specifications, the reform decreases custodial sentence length for both treatment groups, while leaving the probability of imprisonment largely unaffected. The question is which mechanism — severity or reference — can rationalize this pattern.

Table 5: Estimated Impact of the Reform

	Estimated Impact on Mean Sentence
Treatment A <i>(Sentencing Range Downward Shift)</i>	↓ 5 months
Treatment B <i>(Damage Range Upward Shift)</i>	↓ 1 month

Note: Summary of empirical findings on the intensive margin. Both treatment groups experience a decline in mean custodial sentence length.

Treatment B: Evidence of the Reference Channel. Treatment B provides the cleanest test of the reference mechanism. In these categories, statutory sentencing ranges remain unchanged; therefore, under a pure severity interpretation, no change in sentence length should occur. Yet the data show a decline in mean sentence length in both the pooled DD estimates and the dynamic specification.

Because the formal punishment benchmark did not shift, the observed decline cannot be attributed to the severity channel. Instead, it is consistent with Prediction 1: when more severe cases are incorporated into an unchanged statutory range, the reference distribution becomes more severe. Existing cases become relatively mild within their comparison set, leading to lower sentences under reference-dependent evaluation.

The modest magnitude and gradual evolution of the effect in the dynamic specification further supports this interpretation. Unlike a mechanical change in statutory bounds, adjustment through reference updating plausibly requires repeated exposure to the new case mix.

Treatment A: Dominance of the Severity Channel. According to Prediction 2, the observed change reflects the net effect of two opposing forces:

(i) the severity channel, which predicts lower sentences due to a reduced formal punishment benchmark, and

(ii) the reference channel, which predicts higher sentences if cases are repositioned upward within the new distribution.

Empirically, sentence length declines sharply and immediately for Treatment A. The magnitude of the reduction — approximately 5 months relative to a pre-reform mean of 21 months — is economically substantial and highly statistically robust. The dynamic estimates show that the decline occurs without delay and persists over time.

This pattern indicates that the severity channel dominates any countervailing reference effect in Treatment A. The immediate adjustment is consistent with a mechanical response to the downward shift in statutory ranges.

Joint Interpretation. Taken together, the results provide evidence for the coexistence of both mechanisms. The decline in Treatment B identifies the reference channel, as it cannot be explained by severity alone. The larger and immediate decline in Treatment A indicates that the severity channel is operative and quantitatively stronger in categories where statutory ranges changed.

Importantly, the reform does not generate systematic changes in the probability of custodial imprisonment. The effects are confined to the intensive margin, suggesting that statutory ranges primarily anchor punishment severity rather than the binary incarceration decision.

Several limitations merit further discussion. First, in the Czech context, custodial punishment may be substituted for a suspended sentence and other alternative sanctions. If the composition of punishment types shifted around the reform, the intensive-margin estimates could reflect selection effects. Appendix Section A.3 addresses this concern and finds no systematic change in punishment composition.

Second, the pre-reform period is relatively short, as detailed damage reporting began only in 2019. While the dynamic specification shows no evidence of differential pre-trends, future research could examine other sentencing reforms with longer pre-treatment windows to obtain

sharper estimates.

Finally, this paper does not capture the general equilibrium effects - in particular, the response of offenders to these sentencing ranges. A fruitful extension may be to examine this range, testing, for instance, the economic theory of crime (Becker, 1968).

Conclusion

This paper studies how threshold-based statutory categorization shapes discretionary decision-making in criminal sentencing. I exploit a 2020 Czech criminal law reform that shifted nominal damage thresholds for theft offenses, generating plausibly exogenous variation in category membership. The reform created two conceptually distinct sources of variation: (i) categories in which the statutory sentencing range shifted downward (Treatment A), and (ii) categories in which the formal statutory range remained unchanged but the within-category damage distribution expanded upward (Treatment B). This structure allows me to separate a *severity channel*—statutory ranges as signals of intended punishment levels—from a *reference channel*—within-category comparative evaluation as the case mix changes. The paper complements experimental evidence on sentencing mechanisms by showing that severity and reference channels operate in a real-world legal system and can generate spillovers across formally unchanged statutory categories.

Using case-level data and a difference-in-differences design with an external control offense unaffected by the reform, I document three main findings. First, the reform did not generate a robust change in the probability of custodial imprisonment: effects on the extensive margin are small and statistically imprecise. Second, the reform produced economically meaningful reductions in the length of custodial sentences. In Treatment A categories, custodial sentence length falls by roughly 5 months across specifications, consistent with an immediate response to the downward shift in statutory bounds. Third, the length of custodial sentences declines by approximately 1 month in Treatment B, where formal sentencing ranges remain unchanged. Event-study estimates show no evidence of differential pre-trends and indicate that the post-reform effects evolve over time, supporting a causal interpretation.

Taken together, the results are consistent with the coexistence of severity and reference mechanisms. The immediate, sizeable reduction in Treatment A indicates that statutory ranges anchor punishment severity as intended by the legislature. The reduction in Treatment B—despite unchanged statutory bounds—suggests an additional reference-group effect: when a category expands to include more severe cases, previously typical cases become relatively less severe within the category, and punishment levels adjust downward. This implication extends beyond sentencing: in any threshold-based system, changing category boundaries can shift reference groups and induce spillovers even where formal rules do not change.

The findings have two broader implications. Substantively, they indicate that statutory thresholds affect not only the feasible set of punishments but also the comparative environment in which discretionary decisions are made. From a policy perspective, reforms that update thresholds may therefore have indirect behavioral consequences that are not captured by formal range changes alone. Methodologically, the results caution against treating “formally unaffected” categories as clean control groups in threshold-based quasi-experiments: compositional shifts can transmit treatment through reference-group redefinition, biasing conventional comparisons.

Overall, the evidence shows that threshold reforms can reshape discretionary outcomes through both direct severity signals and indirect reference-group shifts, highlighting an underappreciated channel through which threshold-based institutional design influences behavior.

References

- Abrams, D. S., Bertrand, M., & Mullainathan, S. (2012). Do judges vary in their treatment of race? *Quarterly Journal of Economics*, *127*(3), 1327–1355.
- Anderson, J. M., Kling, J. R., & Stith, K. (1999). Measuring interjudge sentencing disparity: Before and after the federal sentencing guidelines. *Journal of Law and Economics*, *42*(S1), 271–308.
- Anwar, S., & Fang, H. (2015). Testing for racial prejudice in the parole board release process: Theory and evidence. *American Economic Review*, *105*(5), 712–716.
- Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of political economy*, *76*(2), 169–217.
- Bjerk, D. (2017). Mandatory minimums and the sentencing of federal drug crimes. *Journal of Legal Studies*, *46*(1), 93–128.
- Bordalo, P., Gennaioli, N., & Shleifer, A. (2012). Saliency theory of choice under risk. *Quarterly Journal of Economics*, *127*(3), 1243–1285.
- Chen, D. L., Moskowitz, T. J., & Shue, K. (2016). Decision making under the gambler's fallacy: Evidence from asylum judges, loan officers, and baseball umpires. *Quarterly Journal of Economics*, *131*(3), 1181–1242.
- Danziger, S., Levav, J., & Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, *108*(17), 6889–6892.
- Drápal, J., & Šoltés, M. (2024). Sentencing decisions around quantity thresholds: Theory and experiment. *Journal of Experimental Criminology*, *20*(4), 1323–1367.
- Englich, B., Mussweiler, T., & Strack, F. (2006). Playing dice with criminal sentences: The influence of irrelevant anchors on experts' judicial decision making. *Personality and Social Psychology Bulletin*, *32*(2), 188–200. Retrieved from <https://doi.org/10.1177/0146167205282152> (PMID: 16382081) doi: 10.1177/0146167205282152
- Kelman, M., Rottenstreich, Y., & Tversky, A. (1996). Context-dependence in legal decision making. *Journal of Legal Studies*, *25*(2), 287–318.
- Kőszegi, B., & Rabin, M. (2006). A model of reference-dependent preferences. *Quarterly Journal of Economics*, *121*(4), 1133–1165.
- Lee, D. S. (2008). Randomized experiments from non-random selection in u.s. house elections. *Journal of Econometrics*, *142*(2), 675–697.
- Leibovitch, A. (2016). Relative judgments. *Journal of Legal Studies*, *45*(2), 281–330.
- Leibovitch, A. (2017). Punishing on a curve. *Northwestern University Law Review*, *111*(5), 1205–1280.

- Saez, E. (2010). Do taxpayers bunch at kink points? *American Economic Journal: Economic Policy*, 2(3), 180–212.
- Skugarevskiy, D. (2017). *Essays in law and economics of enforcement* (Unpublished doctoral dissertation). Graduate Institute of International and Development Studies.
- Ulmer, J. T., & Johnson, B. (2004). Sentencing in context: A multilevel analysis. *Law & Society Review*, 38(4), 777–803.

A Appendix

A.1 Damage report rate

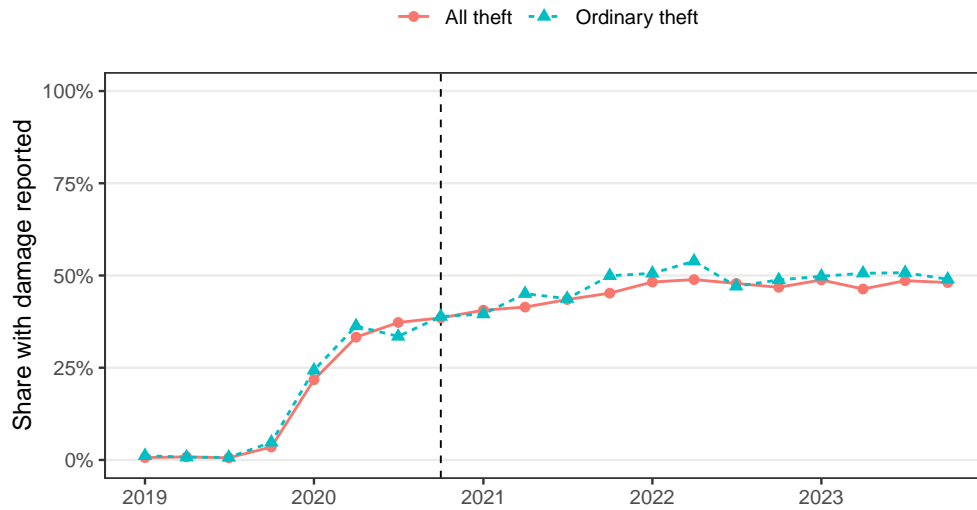


Figure A.1: The rate of cases with damage reported before and after the reform

Note: Dots represent all theft cases; triangles represent ordinary theft cases (cases where damage is the criterion determining the sentencing range that were used as the main sample). The black dashed line denotes the 2020 reform. The date refers to the date on which the sentence takes legal effect, which determines whether the pre- or post-reform legal norm applies.

A.2 Descriptive characteristics of the dataset

Table A.1: Descriptive Statistics of the Main Dataset

	All		Ordinary Cases	
	Before	After	Before	After
Cases	22,383	35,677	936	4,095
Custodial Sentences	14,209	22,885	680	2,954
Suspended Sentences	1,666	2,139	77	323
Mean Custodial Length (m)	14.33	14.12	16.64	16.89
Mean Suspended Length (m)	12.94	12.88	15.26	15.12
Mean Damage (thousand CZK)	68.11	70.38	122.44	159.53
Recidivist Rate (%)	11	12	13	11
Mean Age	32.47	33.42	33.12	33.64
Male (%)	83	85	81	83

Note: The year range is limited to 2019-2023. Ordinary cases are defined as cases with no special circumstances, where the criterion determining the sentencing range is the damage caused. By recidivist, I mean offenders whose previous convictions are counted as aggravating circumstances (under the Criminal Code, it is the court's discretion to consider them as aggravating circumstances).

A.3 Other types of punishment

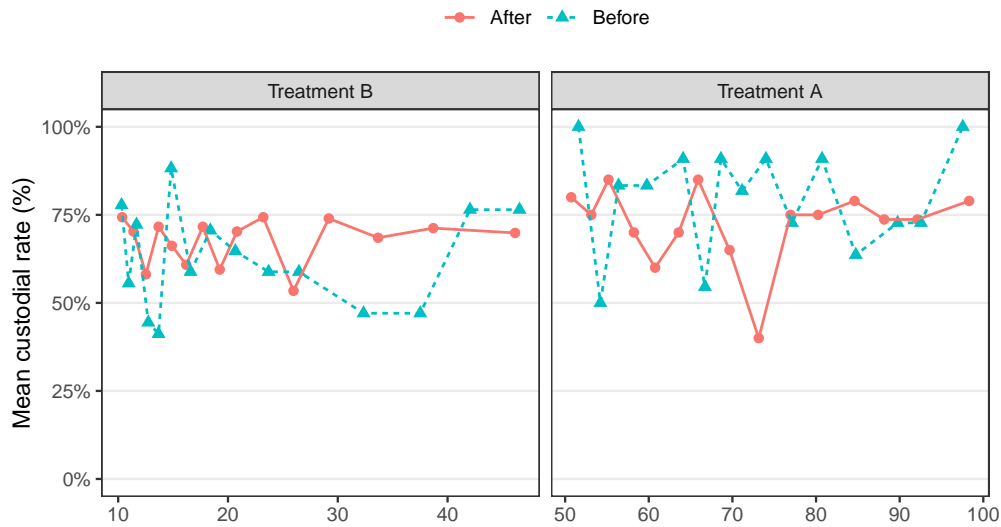
To further investigate the extensive margin of the reform, I estimate a logit model on punishment type. I introduce three categories of punishment: custodial, suspended, and other. Across specifications, the reform does not generate a systematic shift from custodial imprisonment into suspended imprisonment or other sanctions. This mitigates concerns that the intensive-margin effects are driven by compositional changes in punishment type.

Table A.2: Logit estimates of the reform's impact on punishment type

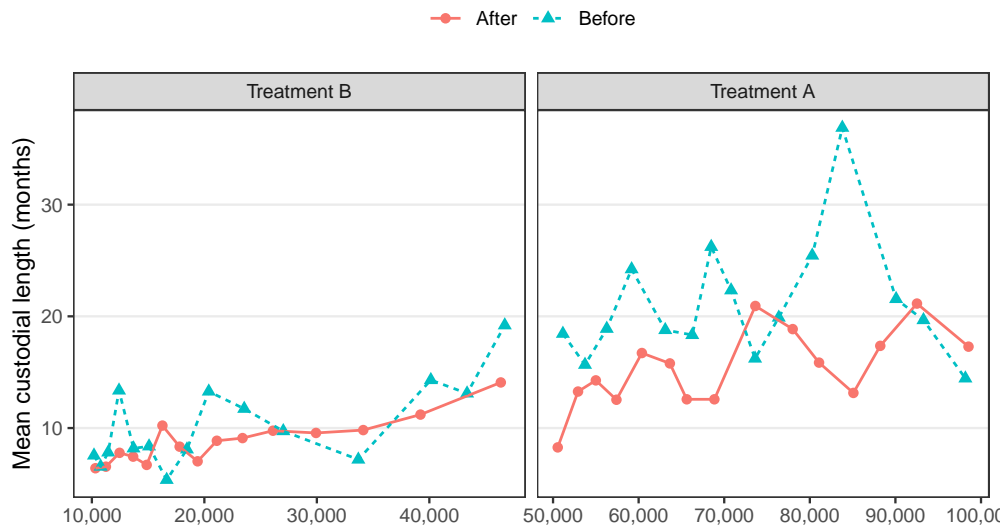
	Suspended		Other
	(1)	(2)	(3)
Treatment A \times After	-0.270 (0.243)	0.184 (0.394)	0.212 (0.351)
Treatment B \times After	0.226 (0.155)	-0.335 (0.284)	-0.284 (0.187)
Num. Obs.	24 520	20 751	20 751
Controls	✓	✓	✓
Judge FE	✓	✓	✓
Quarter FE	✓	✓	✓

A.4 Custodial imprisonment length as a function of damage

Figure A.2 plots custodial sentence length against the monetary damage within the main Treatment A and Treatment B damage intervals, separately for pre- and post-reform cases. The post-reform shift in sentence lengths is not confined to cases close to the statutory thresholds; instead, it is visible across the relevant damage range within each treated group. This supports the interpretation that the baseline DD estimates are not mechanically driven by threshold cases but reflect broader changes in sentencing severity within categories. This supports the interpretation that the baseline DD estimates are not mechanically driven by threshold cases but reflect broader changes in sentencing severity within categories.



(a) custodial sentence rate (extensive margin)



(b) custodial sentence length (intensive margin)

Figure A.2: Effect of the reform for different values of damage

A.5 Alternative control group

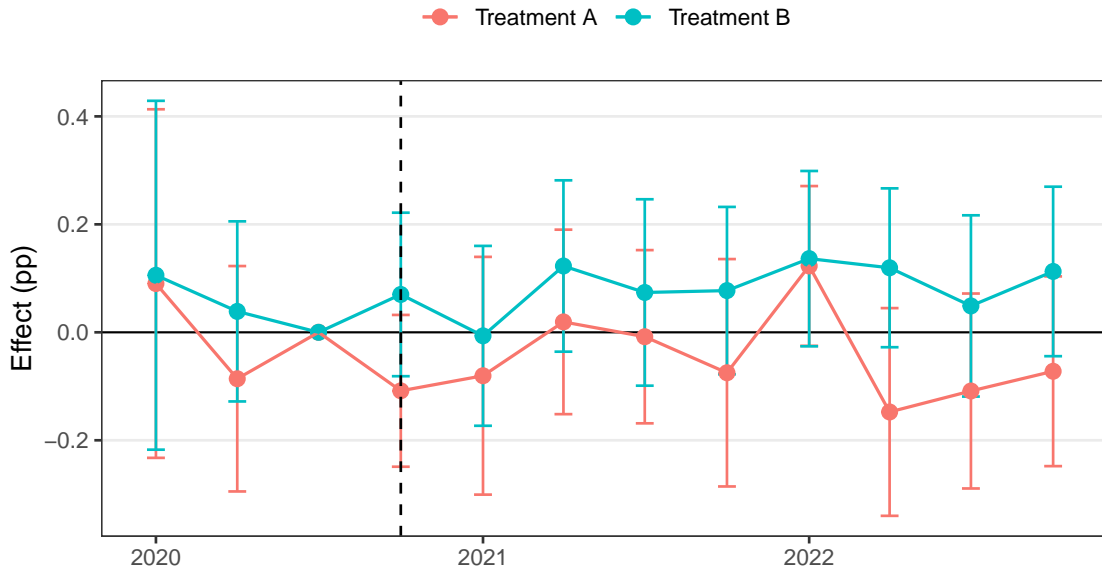
To address the concerns that my findings may be driven by an arbitrary choice of control group, I replicate the results using an alternative control group. I opt for cases under §337 (1) of the Criminal Code (obstruction of an official decision/sentence of banishment), which were not subject to statutory threshold changes.

Overall, the estimates obtained with the alternative control group reinforce the main conclusion that the reform primarily affected the intensive margin of punishment rather than the probability of incarceration.

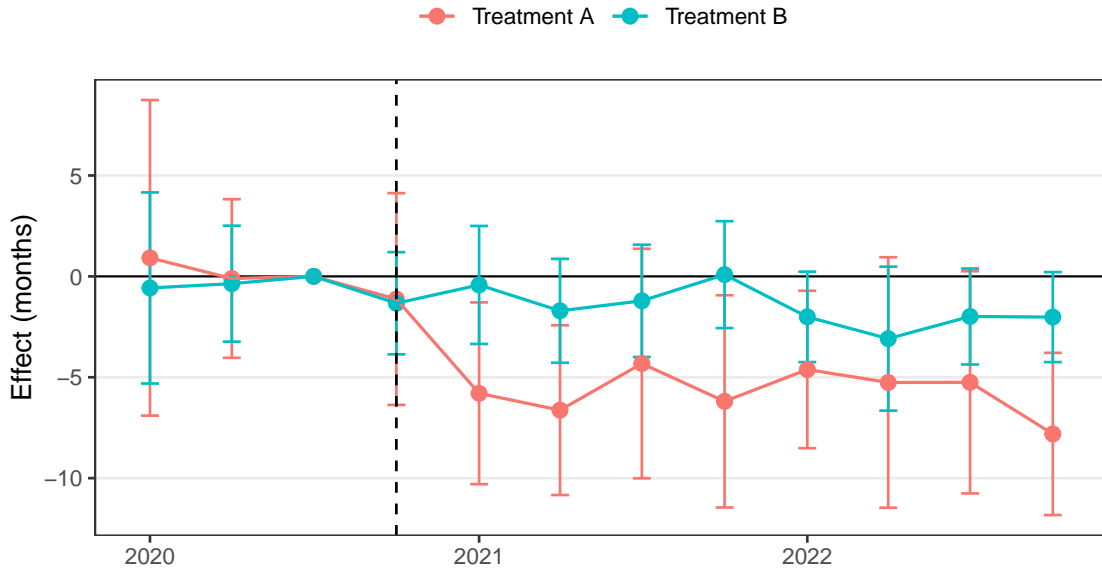
Table A.3: Difference-in-Differences Estimates - Alternative Control Group

	(1)	(2)	(3)	(4)
Panel A. Extensive margin: Custodial imprisonment (LPM)				
Treatment A × After	-0.063 (0.044)	-0.069 (0.043)	-0.041 (0.044)	-0.040 (0.043)
Treatment B × After	0.063+ (0.035)	0.040 (0.037)	0.066+ (0.035)	0.046 (0.037)
Observations	36,971	36,971	36,971	36,971
R^2	0.005	0.070	0.010	0.075
Controls			✓	✓
Judge FE		✓		✓
Quarter FE		✓		✓
Panel B. Intensive margin: Custodial sentence length (months)				
Treatment A × After	-5.893*** (1.470)	-5.516*** (1.441)	-4.868*** (1.335)	-4.455*** (1.316)
Treatment B × After	-1.379+ (0.750)	-1.294+ (0.746)	-1.068+ (0.631)	-0.926 (0.635)
Observations	19,947	19,947	19,947	19,947
R^2	0.057	0.129	0.115	0.183
Controls			✓	✓
Judge FE		✓		✓
Quarter FE		✓		✓

Notes: The table reports pooled difference-in-differences estimates comparing Treatment A and Treatment B cases to control offenses. Standard errors (in parentheses) are clustered at the judge level. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.



(a) quarterly effects on custodial sentence rate (extensive margin)



(b) quarterly effects on custodial sentence length (intensive margin)

Figure A.3: Dynamic DD for custodial sentence - Alternative Control Group

A.6 Alternative samples

As an additional robustness check, I standardize the outcome variables and consider alternative definitions of the treatment groups. Specifically, I construct standardized outcomes using the pre-reform mean and standard deviation of the control group. Both the extensive margin (custodial imprisonment indicator) and the intensive margin (custodial sentence length) are transformed into z-scores prior to estimation.

I also examine broader treatment definitions by pooling theft cases across damage intervals that were affected similarly by the reform. In particular, “Pooled Treatment A” combines theft cases from damage intervals 50k–100k CZK, 500k–1m CZK, and 5m–10m CZK, while “Pooled Treatment B” combines intervals 10k–50k CZK, 100k–500k CZK, and 1m–5m CZK. These pooled groups reflect the fact that the reform shifted statutory thresholds proportionally across several damage categories. For comparison, Panels C and D report results using the original narrow treatment intervals from the baseline specification.

Table A.4 reports the corresponding difference-in-differences estimates. Across specifications, the results are qualitatively consistent with the baseline findings. For the intensive margin, custodial sentence lengths decline following the reform, with statistically significant reductions for Treatment B in the pooled specification and for Treatment A in the original narrow sample. For the extensive margin, the estimated effects remain small and statistically insignificant across all specifications.

Overall, the results indicate that the main findings are robust to both standardizing the outcome variables and redefining treatment groups using broader damage intervals.

Table A.4: Standardized Outcomes and Pooled Treatment Bins

	(1)	(2)	(3)	(4)
Panel A. Extensive margin (standardized): Custodial imprisonment - pooled				
Pooled Treatment A \times After	-0.053 (0.079)	-0.065 (0.077)	-0.026 (0.080)	-0.030 (0.078)
Pooled Treatment B \times After	0.042 (0.052)	-0.019 (0.053)	0.051 (0.050)	-0.003 (0.052)
Observations	38,374	38,374	38,374	38,374
R^2	0.015	0.076	0.020	0.082
Panel B. Intensive margin (standardized): Custodial sentence length - pooled				
Pooled Treatment A \times After	-0.253 (0.233)	-0.150 (0.222)	-0.138 (0.207)	-0.035 (0.203)
Pooled Treatment B \times After	-0.352* (0.160)	-0.341* (0.156)	-0.292* (0.138)	-0.266+ (0.136)
Observations	21,099	21,099	21,099	21,099
R^2	0.203	0.262	0.293	0.346
Panel C. Extensive margin (standardized): Original narrow sample				
Original Treatment A \times After	-0.127 (0.088)	-0.137 (0.087)	-0.082 (0.088)	-0.081 (0.087)
Original Treatment B \times After	0.127+ (0.071)	0.080 (0.074)	0.132+ (0.071)	0.092 (0.074)
Observations	36,971	36,971	36,971	36,971
R^2	0.005	0.070	0.010	0.075
Panel D. Intensive margin (standardized): Original narrow sample				
Original Treatment A \times After	-0.979*** (0.244)	-0.916*** (0.239)	-0.809*** (0.222)	-0.740*** (0.219)
Original Treatment B \times After	-0.229+ (0.125)	-0.215+ (0.124)	-0.177+ (0.105)	-0.154 (0.105)
Observations	19,947	19,947	19,947	19,947
R^2	0.057	0.129	0.115	0.183
Controls			✓	✓
Judge FE		✓		✓
Quarter FE		✓		✓

Notes: The table reports difference-in-differences estimates with standardized outcomes (z-scores). “Pooled” Treatment A and B correspond to pooling damage intervals affected in the same way by the reform; “Original” corresponds to the baseline narrow intervals used in the main text. Standard errors (in parentheses) are clustered at the judge level. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.